1 Overview

Last lecture, we started studying power law degree distributions, which are often observed in networks. These distributions have nice properties such as being a straight line on a log-log plot and having a heavy tail. As a reminder, their formal definition is the following

Definition. *A non-negative random variable X has a power law distribution if there exist constants $c, \alpha > 0$ such that:*

$$Pr(X \geq x) = c \cdot x^{-\alpha}.$$

In the case where X is a discrete random variable, then there exist constants $c', \alpha' > 0$ s.t.:

$$Pr(X = x) = c' \cdot x^{-\alpha'}.$$

2 Power-Laws in Graphs

We first observe that is it impossible to define graphs that follow exactly a power law degree distribution. The reason is that if Δ is the maximal degree in the graph, then $\Pr(X = \Delta + 1) = 0$, which is impossible for a power-law degree distribution. To work around this difficulty, we will consider graphs that approximately follow a power law degree distribution.

2.1 Generating power-law-like graphs

Given c, α, n, we want to generate a degree distribution with power-law-like properties. To circumvent the impossibility observation that we just made, we make the following assumptions:

- The largest degree in a graph is Δ.
- There is exactly one node with degree Δ.

From the second assumption, we infer that

$$\Pr(X = \Delta) = c \cdot \Delta^{-\alpha} = \frac{1}{n}$$

where the first equality is since the degree distribution follows a power law distribution and the second equality is since if we pick a node at random, there is probability $1/n$ that we pick the node with the largest degree. Solving for Δ, we get that $\Delta = (c \cdot n)^{1/\alpha}$.

2.2 Generating random graphs with arbitrarily degree distribution

We got a degree distribution for the graph we want to generate, it remains to generate a graph with a given degree distribution. Consider the following randomized process, known as the *configuration model*:

- We are given a degree distribution $(d_1, \cdots, d_n)$, indicating that node i has degree d_i.
- For each $i \in \{1, \dots, n\}$, create d_i "half-edges" exiting v_i.
- Connect the half edges by choosing a matching uniformly at random, i.e., create disjoint pairs of half-edges uniformly at random.[1]

It is easy to see that after pairing all the half-edges, this simple process outputs a random graph with the desired degree distribution. Note that it may be possible that some nodes will end up with self-loops, and there may be multiple edges between two nodes. But when the average degree of the graph is held constant (independent of n), these are unlikely events.

3 The Rich Get Richer Model

While the configuration model is a useful analytical tool, it does not reveal the emergence of the power law property. Rather, networks generated as the process described above would assume the degree distribution is a power-law and generate a random graph that maintains the prescribed degree distribution.

We therefore wish to give some mathematical explanation as to why so many networks follow a power law degree distribution. These power law degree distributions can be thought as indicators of processes affected by popularity: the choice of whom to connect to in a network indicates the popularity of nodes.

The *rich-get-richer* (or *preferential attachment*) model provides this kind of an explanation. This model is a dynamic process that occurs in discrete time steps $1, \dots, n$. At each time step t, a new node t arrives and generates a directed edge to an existing node (i.e. one of the nodes that arrived in time step $1, \dots, t-1$) according to a probabilistic rule [2]. We can describe the process as follows:

- Nodes arrive in order $1, \dots, n$;
- When node j arrives it generates a link as follows:
  - With probability p, j connects to $i \in \{1, \cdots, j-1\}$ uniformly at random;
  - With probability $1-p$, j chooses $i \in \{1, \cdots, j-1\}$ uniformly at random and connects to the node i connects to.

Figure 1 is an example of this process for $n = 6$, similar to the one we simulated during lecture. Equivalently, the second step of the process described above can be thought of as the following: with probability $1-p$, j connects to $i \in \{1, \cdots, j-1\}$ with probability proportional to is degree.

[1] We do not worry about the case where two half-edges coming out of the same node connect, as these rarely happen, and if they did, the process can just be restarted.

[2] The model can easily be generalized to when a new node connects to k existing nodes, but we will only consider the case $k = 1$ for simplicity.

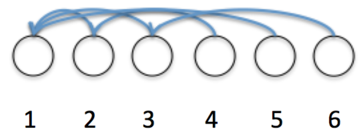

Figure 1: An example of the rich get richer process.

4 Analysis of Rich Get Richer Model

The above model is extremely simple and intuitive, and elegantly captures a rich-get-richer phenomena. But how does this related to power-law graphs? Surprisingly, one can analyze this process and show that the graphs generated according to this process are likely to follow a power law degree distribution. So, our main goal is now to justify this statement, and explain how is it that a process as remarkably simple as the rich get richer model follows a power law distribution. This partially explains why so many networks have a power law distribution since many of them have a preferential attachment type of behavior when new nodes (or persons) join or network.

The randomized process. Let $X_j(t)$ be the random variable denoting the in-degree[3] of node j at time t. The description of the model gives us the following two properties for $X_j(t)$:

- $X_j(t) = 0$ for all $t \leq j$. This is because at time step j, node j has just been introduced; before time step j, node j has not even arrived yet.
- For $t > j$, node j receives an incoming link either due to node t choosing to link to a node uniformly at random (which happens with probability p) *and* the uniformly random decision happens to choose to link to node j (which happens with probability $\frac{1}{t-1}$ as there are $t-1$ nodes in the system at time step t that node t chooses from), *or* node t chooses proportionally to the degree of the nodes (which happens with probability $1-p$) and node j was chosen proportionally to its degree, which happens with probability

  $$\frac{\text{in-degree of } j}{\text{number of links in the graph at step } t}$$

  Thus, the probability of increasing the in-degree of node j (that is, the probability of adding a link from node t to node j) at time step t is

  $$p \cdot \frac{1}{t-1} + (1-p) \cdot \frac{X_j(t-1)}{t-1}.$$

  Looking at this from the lens of time step $t+1$, we have that the probability of increasing the in-degree of node j at time step $t+1$ is

  $$p \cdot \frac{1}{t} + (1-p) \cdot \frac{X_j(t)}{t}.$$

  This is the expression we will make use of in the next step.

[3]The in-degree of a node u in a directed graph is the number of edges with endpoint u.

A deterministic approximation. To simplify the analysis, we consider a deterministic approximation of the process above. The benefits of this approximation is that we deal with a smooth and deterministic process, instead of a randomized process with discrete "jumps". Denote $x_j(t)$ to be the continuous variable that is analogue to $X_j(t)$ and that will be used for the approximation. Observe that the probability of adding a link (and thus the probability of increasing the in-degree) can be thought as the rate of change of $X_j(t)$, which is the derivative in the continuous case. By analogy to the analysis above, we get

- $x_j(j) = 0$.
- $\frac{dx_j}{dt} = p \cdot \frac{1}{t} + (1-p) \cdot \frac{x_j}{t}$ for $t > j$.

Let $q = 1 - p$, then,

$$\frac{dx_j}{dt} = p \cdot \frac{1}{t} + q \cdot \frac{x_j}{t} = \frac{p + qx_j}{t}$$

To solve this differential equation, we integrate and get:

$$\begin{aligned}
\int \frac{1}{p + qx_j} \frac{dx_j}{dt} dt &= \int \frac{1}{t} dt \\
\log(p + qx_j) &= q \log t + c \\
e^{\log(p+qx_j)} &= e^{q \log t + c} \\
p + qx_j &= At^q
\end{aligned}$$

where c is a constant and $A = e^c$. We therefore get

$$x_j = \frac{At^q - p}{q}.$$

We then use the initial condition to find A:

$$x_j(j) = \frac{Aj^q - p}{q} = 0.$$

Solving this, we find that $A = pj^{-q}$ and we get that

$$x_j(t) = \frac{pj^{-q}t^q - p}{q} = \frac{p}{q}\left(\left(\frac{t}{j}\right)^q - 1\right).$$

To obtain the degree distribution, we wish to get the fraction of nodes with degree at least k given some time t. The nodes that satisfy $x_j(t) \geq k$ are the nodes j such that $\frac{p}{q}\left(\left(\frac{t}{j}\right)^q - 1\right) \geq k$, or equivalently (by solving for j), the nodes j such that $j \leq t\left(\left(\frac{q}{p}\right)k + 1\right)^{-1/q}$. The degrees at time step t are $x_1(t), \cdots, x_t(t)$, so we get that the fraction of nodes with degree at least k is

$$\left(\left(\frac{q}{p}\right)k + 1\right)^{-1/q} \approx c \cdot k^{-\alpha}$$

with $c \approx q/p$ and $\alpha = 1/q$.